DNN-Based Speech Synthesis for Arabic: Modelling and Evaluation

✉Amal Houidhek[1,2], Vincent Colotte[2], Zied Mnasri[1], and Denis Jouvet[2]

[1] Electrical Engineering Department, Ecole Nationale d'Ingénieurs de Tunis, University Tunis El Manar, Tunisia
[2] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{amal.houidhek,vincent.colotte,denis.jouvet}@loria.fr,
zied.mnasri@enit.utm.tn

Abstract. This paper investigates the use of deep neural networks (DNN) for Arabic speech synthesis. In parametric speech synthesis, whether HMM-based or DNN-based, each speech segment is described with a set of contextual features. These contextual features correspond to linguistic, phonetic and prosodic information that may affect the pronunciation of the segments. Gemination and vowel quantity (short vowel vs. long vowel) are two particular and important phenomena in Arabic language. Hence, it is worth investigating if those phenomena must be handled by using specific speech units, or if their specification in the contextual features is enough. Consequently four modelling approaches are evaluated by considering geminated consonants (respectively long vowels) either as fully-fledged phoneme units or as the same phoneme as their simple (respectively short) counterparts. Although no significant difference has been observed in previous studies relying on HMM-based modelling, this paper examines these modelling variants in the framework of DNN-based speech synthesis. Listening tests are conducted to evaluate the four modelling approaches, and to assess the performance of DNN-based Arabic speech synthesis with respect to previous HMM-based approach.

Keywords: Parametric speech synthesis · Hidden Markov Models · Decision tree· Deep neural network · Arabic language.

1 Introduction

Statistical parametric speech synthesis (SPSS) approach has been widely used in the last decade. It presents the advantages of being trainable and making possible changing voice characteristics [4]. SPSS is based on Hidden Markov Models to model speech parameters, as in HTS toolkit (Hidden Markov Models speech synthesis system). HTS has been applied to many languages e.g., English [18], Japanese [24] and Arabic [1] and produces speech of rather good quality. HTS requires the description of each speech segment with a set of contextual features that comprises all factors affecting the pronunciation of the corresponding sound (e.g., linguistic, prosodic, phonological information). A standard set of around

50 features was suggested in [18]. Part of the features are language dependent, therefore some modifications of the features set was suggested in [13] and [14] (either ignoring or adding information) to be adapted to the specificities of respectively German and French languages. Actually, the choice of contextual features is primordial as it affects the speech quality.

Arabic speech synthesis using HTS was initiated in [1]; the conventional system was adapted to Arabic with a modification of the excitation model and speech parameters to enhance the speech quality. Later, STRAIGHT vocoder [11] was used in [12] to generate a higher-quality Arabic speech. [7] focused on phonological particularities of Modern Standard Arabic (MSA) [2]. Two phenomena were highlighted, namely gemination [16] (i.e. a geminated consonant is twice as long as its simple counterpart) and vowel quantity (short vowel vs. long vowel) [17] (i.e. a long vowel is twice as long as its short counterpart). In [7] subjective and objective evaluations showed that considering the geminated consonants (resp long vowels) as fully-fledged phonemes or as the same phonemes as their simple (resp short) counterparts leads to similar speech quality as long as the information about gemination and vowel quantity are included in the set of contextual features.

According to [23, 4], the naturalness of HTS output speech has never reached the level of unit-selection-generated speech [8]. This is due to three major reasons; vocoding, inaccurate acoustic model and over-smoothing. In SPSS, acoustic models match the contextual features to the corresponding speech parameters. In this approach, the mapping from contextual features to speech parameters is achieved based on decisions trees [10], which are described as shallow architectures, therefore, they are judged inefficient to represent complex dependencies between contextual features and acoustic parameters. Though temporal-domain oversmoothing has almost no effect on quality, frequency-domain oversmoothing is mainly due to the training algorithm accuracy, and may degrade the quality of output speech by causing an envelope effect [25].

To cope with these issues, previous works suggested replacing decision trees by DNN [22] or using external models for duration [21]. Results showed that DNN outperformed HMM in terms of speech quality and naturalness of produced speech for English language [23, 19]. This paper aims at introducing DNN in parametric speech synthesis for Arabic and investigating if DNN benefit from the explicit differentiation of different phoneme classes unlike HMM [7]. The paper is organised as follows. Section 2 presents various choices of speech unit modelling for HMM-based speech synthesis in Arabic. Section 3 details DNN-based speech synthesis. Section 4 compares and discusses the various speech unit modelling approaches. Finally, Section 5 presents the evaluations of the HMM-based and DNN-based approaches for Arabic speech synthesis.

2 Speech Unit Modelling for Arabic

One of the Arabic speech modelling issues is how should gemination and vowel quantity be regarded: whether is it enough to add gemination and vowel quantity

information to the features set, or is it better to consider a geminated consonant (resp. a long vowel) as fully-fledged speech unit in the modelling?

2.1 Speech Unit Modelling

This problem has been dealt with in [7] for HMM-based Arabic speech synthesis, where four modelling approaches are proposed; differentiating geminated consonants (resp long vowels) from simple consonants (resp short vowels) or merging them:

- ***C2V2***: This is the most detailed model, where a simple consonant (e.g., /d/) and its geminated counterpart (e.g., /dd/) are modelled by two different units. In the same way, short vowels (e.g., /a/) and their long counterparts (e.g., /aa/) have distinct models.
- ***C1V1***: It is the most compact model, where geminated and simple consonants are modelled with the same unit, as well for vowels, long and short vowels are modelled with the same unit.
- ***C1V2***: In this approach, a single unit models both a geminated consonant and its simple counterpart, whereas a long vowel and its short counterpart are modelled by two different units.
- ***C2V1***: This approach uses a single unit to model both a long vowel and its short counterpart. Whereas for consonants, two units are used, one for the simple consonant and one for its geminated counterpart.

Note that in all cases, gemination and vowel quantities characteristics are included into the set of contextual features.

2.2 Experiments with HMM-Based Modelling

This section summarizes the experiments described in [7], which were conducted to compare the four modelling approaches listed above in the framework of HMM-based synthesizer. The speech data used to train the speaker-dependent models with HTS was extracted from the corpus developed in [6]. The training set consists of 1565 utterances recorded by a male-speaker at 48 KHz sampling rate, whereas the test set comprises 30 utterances. Subjective evaluations showed that the four modelling approaches lead to similar speech quality and present almost the same degree of degradation when compared to the natural speech [7]. Moreover, a one-to-one comparison of the four models showed that listeners had no clear preference for a particular one.

Consequently, differentiating geminated consonants (resp. long vowels) from simple consonants (resp. short vowels) or merging them lead to a similar speech synthesis quality. HMM-based speech synthesis did not benefit from the explicit differentiation between the different classes of phonemes (i.e., simple vs. geminated consonants and short vs. long vowels).

3 DNN-Based Speech Synthesis

3.1 DNN vs. Decision Trees

Decision trees used in HMM-based speech synthesis, present major shortcomings [23, 19]. They are inefficient to model complex functions and dependencies between contextual features and acoustic parameters. Since the set of contextual features contains around 50 features, it requires large decision trees to be modelled. Besides, during the training, decision trees split the training data into sub-clusters and use different parameters for each cluster [22]. This process affects the clustering of the context-dependent distributions, thus the estimation of the distributions for speech parameters prediction. According to [3], DNN are able to represent complicated functions, besides, the weights of DNN are trained from all the training data.

3.2 DNN-Based Speech Synthesis System

In DNN-based speech synthesis, the contextual features are mapped to the output vector, which contains spectral and excitation parameters and their dynamic features. Weights of the DNN are trained using pairs of input and output features extracted from training data to minimize the error between the mapped output predicted from the given input and the target output. Finally, a vocoder is used to process the generated speech parameters to produce a speech signal.

3.3 Merlin Toolkit

Merlin speech synthesis toolkit for neural network-based speech synthesis was introduced in [20]. Merlin proposes a variety of architecture e.g., a standard feed-forward neural network, recurrent neural network (RNN) and long short-term memory (LSTM). Moreover, Merlin supports WORLD [15] and STRAIGHT [11] vocoder. The input vector of the neural network includes numerical values (e.g., the number of phonemes in the syllable, position of the syllable in the word...) and binary answers to questions about identities of the phonemes context (e.g., is the current phoneme "a"...) and other characteristics.

4 Evaluation of Speech Unit Modelling

4.1 Experiment Conditions

The evaluation of the speech unit modelling approaches (C2V2, C1V1, C1V2 and C2V1) is conducted using the training and test sets described in Section 2. The contextual features are the same as in [7]. The input vector consists of 816 features where 771 of them are binary answers to questions about context of the phonemes (e.g., identity of the phoneme, identity of the vowel of the current syllable...), whereas the remaining 45 are numeric values (e.g., position of the phoneme in the syllable, the duration of the phoneme and of the state in frames, frame position

within the state and the phoneme, the state position within the phoneme forward and backward etc.). Several tests were conducted to choose the DNN architecture that can generate the best speech quality. In current experiments, the DNN is composed of 4 layers of 1024 units with tanh transfer function plus one BLSTM (bidirectional LSTM) on the upper layer with 512 units to consider the sequential aspect of the speech [5]. During the experiment, WORLD vocoder is used to extract 60-dimensional MCCs (Mel-Cepstral Coefficients), 5-dimensional BAPs (Band APeriodicities) and log (F0) at a frame length of 5 ms.

4.2 Objective Evaluation of Duration

An objective evaluation is conducted with respect to duration of sounds. For speech signals produced with each modelling approach (C2V2, C1V1, C1V2 and C2V1) the average, over the vowels, of the ratios between the mean duration of long vowels (LV) and the mean duration of corresponding short vowels (SV) is calculated as well as the average ratio for geminated consonants (GC) vs. simple consonants (SC). Only phonemes with more than 10 occurrences for each class (simple/geminated consonants and short/long vowels) are considered. The calculated average ratios are compared to those obtained on natural speech.

Table 1. Duration ratios.

	LV / SV	GC / SC
Number of occurrence	262 / 884	104 / 1315
C2V2	1.7	2.1
C1V1	1.7	2.1
C1V2	1.7	2.1
C2V1	1.8	2.2
Natural	2.0	2.1

Values in Table 1, show that for the four modelling approaches, the ratios between the predicted durations of long vowels (LV) and short vowels (SV) are lower than those calculated for natural speech. However, the ratios between predicted durations of geminated consonants (GC) and predicted durations of simple consonants (SC) are similar to those calculated on natural speech.

Root mean square error (RMSE) between natural duration and predicted durations was calculated on the different phoneme classes (simple and geminated consonants, and short and long vowels). Values of RMSE are presented in Fig. 1. Results show that for each class, the C2V2 model (the most detailed model) leads to lower RMSE than the other approaches (C1V1, C1V2 and C2V1).

Normalized root mean square error (NRMSE) is calculated by considering the mean duration values of each phoneme class (NRMSE=RMSE/M, where M is the mean duration). The obtained results are presented in Fig. 2. NRMSE of the model C2V2 presents a significant decrease. Meanwhile, for each phoneme class, the other approaches present similar values of NRMSE.

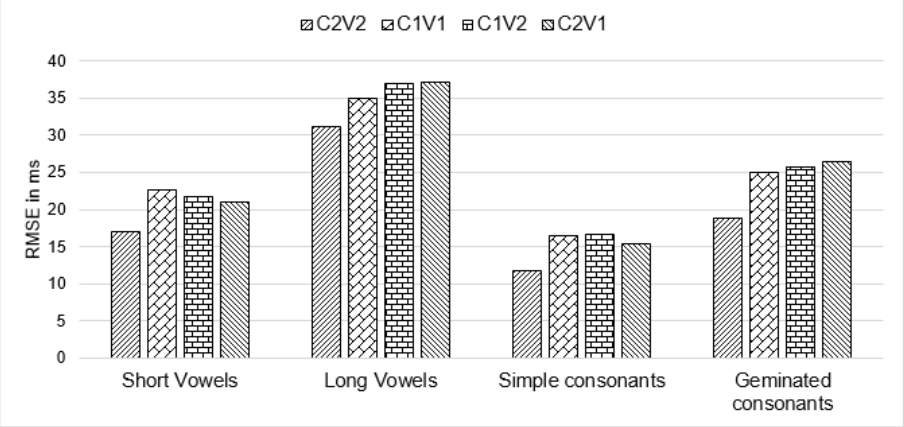

Fig. 1. RMSE between natural and predicted durations

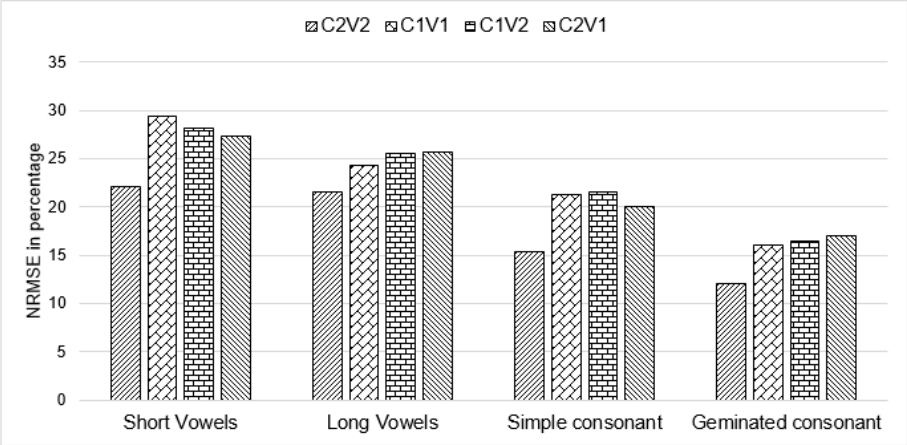

Fig. 2. NRMSE between natural and predicted durations

4.3 Comparison of Modelling Approaches

A preference test [9] was conducted to compare the four proposed approaches. 18 Arabic native speakers participated in this evaluation. Each one evaluated a set of 20 pairs of speech signals; each pair consists of the same utterance produced with two different approaches. The order of presentation of the speech signals is randomly chosen for each trial. During the evaluation, participants were asked to point to the preferred signal based on the global quality of produced speech by answering the following question: *"How do you judge the quality of the second signal compared to the first one?"* and giving a score from 1 to 7 ranging from much worse to much better. Results of comparison are shown in Fig. 3. To analyse the results, scores were grouped to get three possible rates; first preferred (scores 1 and 2 corresponding to much worse and worse) , no preference (scores 3, 4 and 5 corresponding to a little worse, about the same and a little better) and second preferred (scores 6 and 7 corresponding to better and much better) :

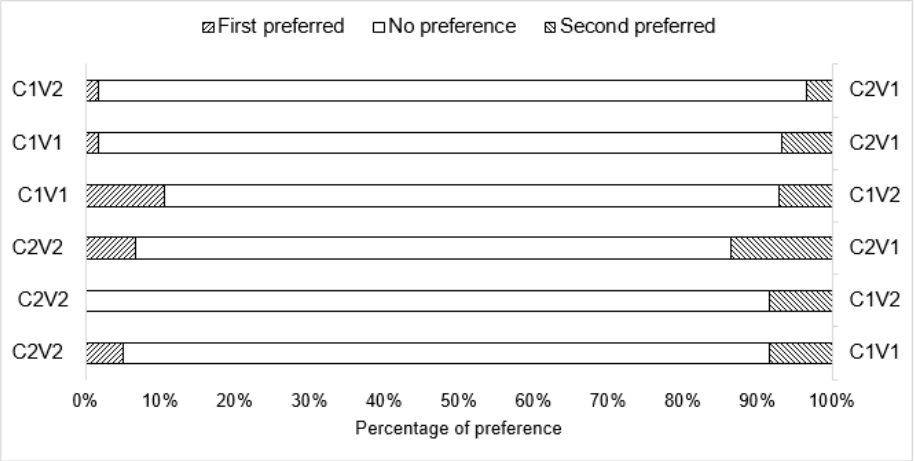

Fig. 3. Results of preference test

The one-to-one comparison shows that listeners had no clear preference for one particular approach. Although, C2V2 leads to a better prediction of duration, the listening tests show that differentiating geminated consonants (resp. long vowels) from simple consonants (resp. short vowels) or merging them leads to similar speech synthesis quality.

5 Evaluation of DNN-Based Speech Synthesis

5.1 Experiments Conditions

DNN-based speech synthesis was evaluated through a comparison to the standard HMM-based speech synthesis (based on decision trees) and to speech processed

by the WORLD vocoder. Evaluation data consists of 30 stimuli generated using context-dependent HMM and the model C2V2, 30 stimuli produced using DNN and the model C2V2 and 30 stimuli processed by copy synthesis i.e., natural signals were analysed using the vocoder WORLD, then they were reconstructed based on the extracted speech parameters using the same vocoder WORLD. Note that participants are Arabic native speakers and they are neither specialists in phonetics nor accustomed to speech evaluation.

5.2 Evaluation of Global Quality

MOS (Mean Opinion Score) tests [9] were conducted to assess the global quality and naturalness of produced speech signals. The global quality refers to the overall quality of generated signals. The naturalness is assessed based on the intonation and the rhythm of synthesized speech signals. 15 listeners participated in these tests. Each one evaluated a set of 20 stimuli i.e., 10 from each set (stimuli produced by HMM and DNN-based speech synthesis systems) and judged the corresponding overall quality and naturalness. Listeners were asked to answer the following question: "*In terms of general impression, how do you judge the overall quality and the naturalness of what you have just heard?*" and give a score from 1 to 5 ranging from very bad to excellent. Fig. 4 shows the MOS scores and the associated 95% confidence interval. Results show that signals produced with DNN-based speech synthesis, have higher MOS scores in terms of overall quality and naturalness than those generated with HMM-based speech synthesis system.

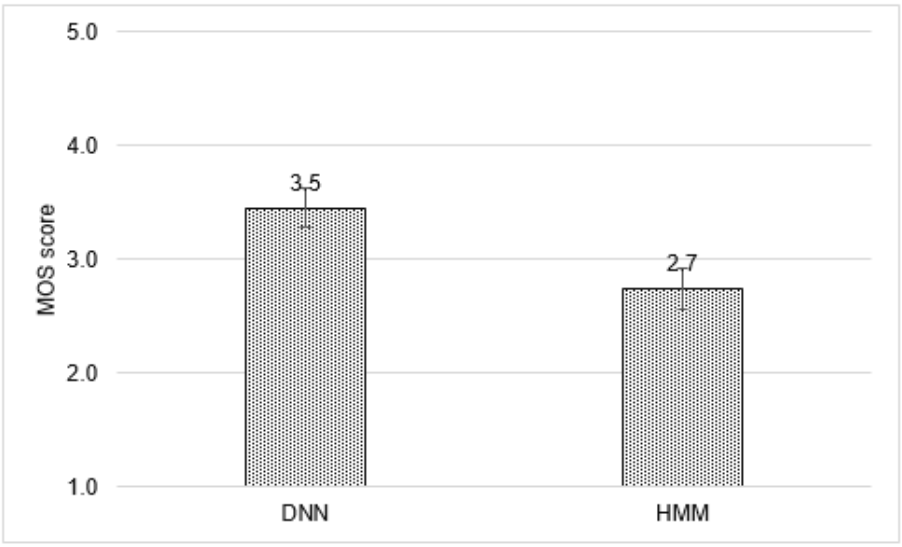

Fig. 4. Results of global quality evaluation.

5.3 Evaluation of Degradation

DMOS (Degradation Mean Opinion Score) tests [9] were conducted to evaluate the degree of degradation caused by the used toolkits HTS (for HMM-based speech synthesis) and Merlin (for DNN-based speech synthesis). Speech signals from each set are compared to the natural speech. Nine listeners participated in these tests, each one evaluated a set of 30 pairs, where each pair consists of the same utterance produced by DNN, HMM-based speech synthesis systems or copy-synthesis and the corresponding natural signal.

Note that the reference (natural signal) is always presented first. Participants evaluated the degradation of signals by answering the following question: *"How do you judge the degradation of the second signal compared to the first one?"*, based on the five-point degradation category scale ranging from very annoying degradation to inaudible degradation. The obtained results are presented in Fig. 5 with the associated 95% confidence interval. The higher the score is, the lower the degradation is.

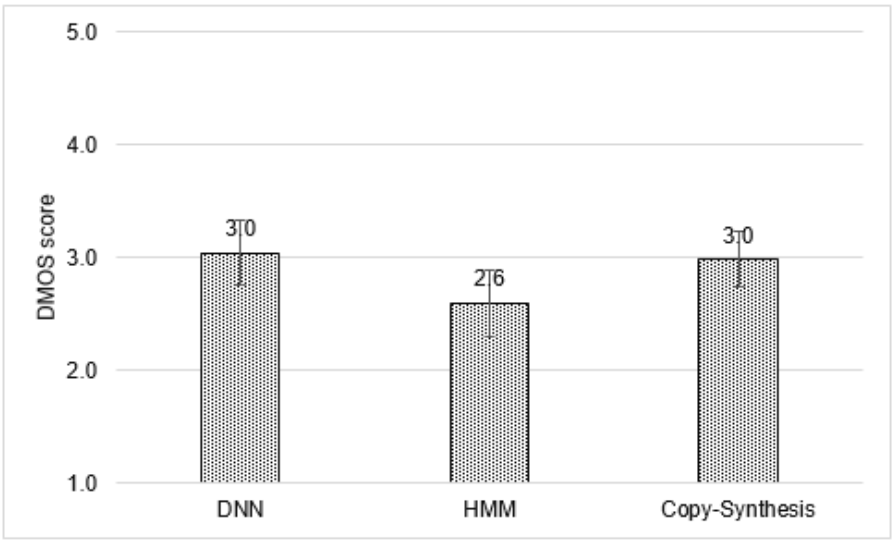

Fig. 5. Results of degradation evaluation.

Results show that the degree of degradation obtained with Merlin is similar to the one obtained by copy-synthesis, and lower than the one obtained with HMM-based speech synthesis system.

5.4 Comparison of DNN and HMM Performance

A preference test [9] was conducted to compare the performance of HMM to DNN-based speech synthesis approaches. Signals generated by Copy-synthesis

using the vocoder WORLD were included in this test as well. The comparison was established with respect to the quality of produced speech. Stimuli are compared to each other. 18 listeners participated in this evaluation. Each one evaluated a set of 30 pairs of speech signals; each pair consists of the same utterance produced with two different approaches. The order of presenting the speech signals is randomly chosen for each trial. Participants were asked to point to the preferred signal based on the global quality of produced speech, by answering this question *"How do you judge the quality of the second signal compared to the first one?"* and giving a score from 1 to 7 ranging from much worse to much better. Scores were grouped in the same way like for Fig. 3.

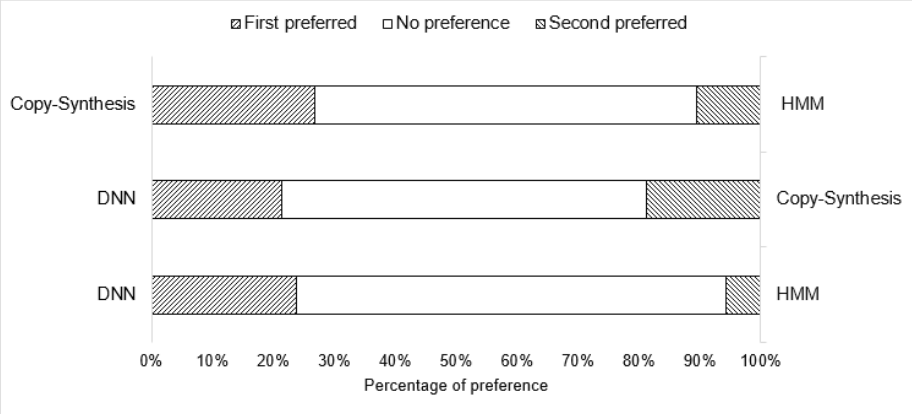

Fig. 6. Results of preference test

Comparison results in Fig. 6 show that signals produced with DNN-based approach and copy-synthesis are preferred when compared to signals produced by HMM-based approach. This is consistent with the results on the evaluation of the global quality: the use of deep neural networks to map the contextual features to the corresponding acoustic parameters is more efficient than the mapping achieved with the decision trees as used in HMM-based speech synthesis system.

6 Conclusions

This paper studied the use of deep neural network in Arabic speech synthesis. Both HMM and DNN-based speech synthesis require the qualification of each text segments with a set of contextual features that comprise all factors (e.g., linguistic, prosodic, phonological...) affecting the pronunciation of the corresponding sound. Part of the set is language dependent, therefore, for Arabic language, two phonological phenomena are highlighted, namely gemination and vowel quantity (short/long). Two extra features are added to the set of contextual features to take into account those specificities.

A variety of possible modelling approaches of speech segments have been investigated such as, the use of different units for modelling long vs. short vowels, and/or the use of different units for modelling simple vs. geminated consonants. These combinations have been compared to another one, where a short vowel and its long counterpart are modelled with the same unit, and a geminated consonant and its simple counterpart are modelled with the same unit. Subjective evaluation of the four speech unit modelling approaches (C2V2, C1V1, C1V2 and C2V1) using Merlin showed that they lead to similar speech quality. However, a better prediction of duration is obtained when using the C2V2 approach (the most detailed). This model attained the lowest RMSE compared to the other models (C1V1, C1V2 and C2V1). Thus, DNN has been more successful to take advantage of specificities of Arabic language.

The second part of this paper focused on assessing the performance of DNN in Arabic speech synthesis. DNN provides an efficient mapping from contextual features to acoustic parameters. This is confirmed by the results of subjective evaluations, which showed that the use of a deep neural architecture in speech synthesis (more specifically in predicting the speech parameters) enhanced the accuracy of acoustic modelling so that the quality of DNN-generated speech is better than the one of HMM-based speech synthesis for Arabic language.

Acknowledgements. This research work was conducted under PHC-Utique Program in the framework of CMCU (Comité Mixte de Coopération Universitaire) grant N 15G1405.

References

1. Abdel-Hamid, O., Abdou, S.M., Rashwan, M.: Improving arabic hmm based speech synthesis quality. In: 9th International Conference on Spoken Language Processing. INTERSPEECH'06. Pittsburgh, Pennsylvania (2006)
2. Al-Ani, S.H.: Arabic phonology: An acoustical and physiological investigation, vol. 61. Walter de Gruyter (1970)
3. Bengio, Y., et al.: Learning deep architectures for ai. Foundations and trends® in Machine Learning 2(1), 1–127 (2009)
4. Black, A.W., Zen, H., Tokuda, K.: Statistical parametric speech synthesis. In: International Conference on Acoustics, Speech and Signal Processing. ICASSP'07. vol. 4, pp. IV–1229. IEEE (2007)
5. Fan, Y., Qian, Y., Xie, F.L., Soong, F.K.: TTS synthesis with bidirectional LSTM based recurrent neural networks. In: 15th Annual Conference of the International Speech Communication Association. Singapore (2014)
6. Halabi, N.: Modern Standard Arabic Speech Corpus. Ph.D. thesis, University of Southampton (2015)
7. Houidhek, A., Colotte, V., Mnasri, Z., Jouvet, D., Zangar, I.: Statistical modelling of speech units in HMM-based speech synthesis for arabic. In: LTC'17-8th Language & Technology Conference. Poznan, Poland (2017)
8. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: International Conference on Acoustics, Speech,

and Signal Processing. ICASSP'96. Conference Proceedings. vol. 1, pp. 373–376. IEEE, Atlanta, GA, USA (1996)
9. ITU: 800, methods for subjective determination of transmission quality. International Telecommunication Union (1996)
10. Jurafsky, D.: Speech and language processing: An introduction to natural language processing. Computational linguistics, and speech recognition (2000)
11. Kawahara, H., Masuda-Katsuse, I., De Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech communication 27(3), 187–207 (1999)
12. Khalil, K.M., Adnan, C.: Arabic HMM-based speech synthesis. In: International Conference on Electrical Engineering and Software Applications. ICEESA'13. pp. 1–5. IEEE, Hammamet, Tunisia (2013)
13. Krstulovic, S., Hunecke, A., Schröder, M.: An HMM-based speech synthesis system applied to german and its adaptation to a limited set of expressive football announcements. In: 8th Annual Conference of the International Speech Communication Association. pp. 1897–1900. Citeseer, Antwerp, Belgium (2007)
14. Maguer, S.L., Barbot, N., Boeffard, O.: Evaluation of contextual descriptors for HMM-based speech synthesis in french. In: 8th Workshop on Speech Synthesis. Barcelona, Spain (2013)
15. Morise, M., Yokomori, F., Ozawa, K.: World: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE transactions on Information and Systems 99(7), 1877–1884 (2016)
16. Newman, D.: The phonetic status of arabic within the world's languages: the uniqueness of the lughat al-daad. Antwerp papers in linguistics. 100, 65–75 (2002)
17. Selouani, S.A., Caelen, J.: Arabic phonetic features recognition using modular connectionist architectures. In: IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications. IVTTA'98. Proceedings. pp. 155–160. IEEE, Torino, Italy (1998)
18. Tokuda, K., Zen, H., Black, A.W.: An HMM-based speech synthesis system applied to english. In: IEEE Speech Synthesis Workshop. pp. 227–230. Santa Monica, CA, USA (2002)
19. Watts, O., Henter, G.E., Merritt, T., Wu, Z., King, S.: From hmms to dnns: where do the improvements come from? In: International Conference on Acoustics, Speech and Signal Processing. ICASSP'16. pp. 5505–5509. IEEE, Lujiazui (2016)
20. Wu, Z., Watts, O., King, S.: Merlin: An open source neural network speech synthesis system. Proc. SSW, Sunnyvale, USA (2016)
21. Zangar, I., Mnasri, Z., Colotte, V., Jouvet, D., Houidhek, A.: Duration modeling using DNN for arabic speech synthesis. In: 9th International Conference on Speech Prosody. pp. 597–601. Poznan, Poland (2018)
22. Zen, H.: Deep learning in speech synthesis. In: SSW. p. 309. Barcelona, Spain (2013)
23. Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: International Conference on Acoustics, Speech and Signal Processing. ICASSP'13. pp. 7962–7966. IEEE (2013)
24. Zen, H., Tokuda, K., Masuko, T., Kobayasih, T., Kitamura, T.: A hidden semi-markov model-based speech synthesis system. IEICE transactions on information and systems 90(5), 825–834 (2007)
25. Zhang, M., Tao, J., Jia, H., Wang, X.: Improving hmm based speech synthesis by reducing over-smoothing problems. In: 6th International Symposium on Chinese Spoken Language Processing. ISCSLP'08. pp. 1–4. IEEE, Kunming, China (2008)